Hugh O'Connor

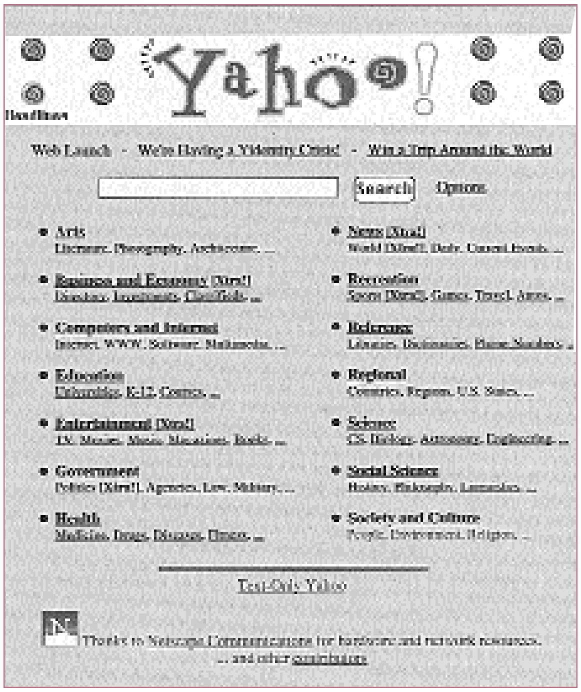
WEBCRAWLER™
LIGHTNING FAST WEB SEARCH

Searching the Internet Through the World Wide Web

The Internet is an enormous, dynamic, and evolving multimedia encyclopedia of all human knowledge and opinion. You can tap this searchable encyclopedia from a desktop computer. Within the last several years great progress has been made in making the Internet's store of knowledge and human expression more precisely accessible; but we are not yet at the point where precise subject searching of the Internet is generally easy, common, consistent, or well-understood. The Net is, we may say, still in its "Wild West" phase.

This article will cite and briefly discuss some of the major Internet search and retrieval tools, and lists currently available on the World Wide Web (WWW or the Web). The Web is the multimedia portion of the Internet containing images, text files, and audio and video clips. This article's approach will be to concentrate on tools that, in the author's experience, are the most comprehensive, powerful, and useful for scanning large regions of the Internet. Many good systems will go unmentioned in this review; but I have attempted to keep the list and my comments short in the belief that it is better to know about a few really good Internet searching tools than about the many, many good but limited and overlapping ones that are constantly appearing.

The search screen of Yahoo! which may be found at http://www.yahoo.



Headlines

Web Launch - We're Having a Yidentity Crisis! - Win a Trip Around the World

Search Options

- Arts
  Literature, Photography, Architecture, ...
- News [Xtra!]
  World [Xtra!], Daily, Current Events, ...
- Business and Economy [Xtra!]
  Directory, Investments, Classifieds, ...
- Recreation
  Sports [Xtra!], Games, Travel, Autos, ...
- Computers and Internet
  Internet, WWW, Software, Multimedia, ...
- Reference
  Libraries, Dictionaries, Phone Numbers, ...
- Education
  Universities, K-12, Courses, ...
- Regional
  Countries, Regions, U.S. States, ...
- Entertainment [Xtra!]
  TV, Movies, Music, Magazines, Books, ...
- Science
  CS, Biology, Astronomy, Engineering, ...
- Government
  Politics [Xtra!], Agencies, Law, Military, ...
- Social Science
  History, Philosophy, Linguistics, ...
- Health
  Medicine, Drugs, Diseases, Fitness, ...
- Society and Culture
  People, Environment, Religion, ...

Text-Only Yahoo

Thanks to Netscape Communications for hardware and network resources.
... and other contributors

What Can Be Searched on the Internet?

As of this writing, the Internet has grown to include:

- Several thousand telnet and FTP sites, containing an estimated 3 million accessible files, including duplicates and alternate versions. Telnet is a program that connects your computer to another one, and allows you to search the remote computer as if your computer were a linked (connected) terminal. FTP is the Internet's "file transfer protocol," a "downloading" utility.
- About 9,000 gophers (menu-based information search systems).
- Over 14,000 USENET newsgroups (online discussion groups whose messages may be accessed via the Internet).
- Almost 13,000 email-based mailing lists (online discussion groups whose messages appear in your email box once you subscribe) running on the listserv, listproc or majordomo list management software. These discussion groups are sometimes generically called listservs.
- More than 5 million World Wide Web pages.
- Several hundred substantial, searchable databases-like the online U.S. Code and Code of Federal Regulations-that are popping up on the Internet every day, faster than the most dedicated Internet students can track. Many of these are based on the WAIS (Wide-Area Information Server) system.

What are the Best Known Internet Search Tools?

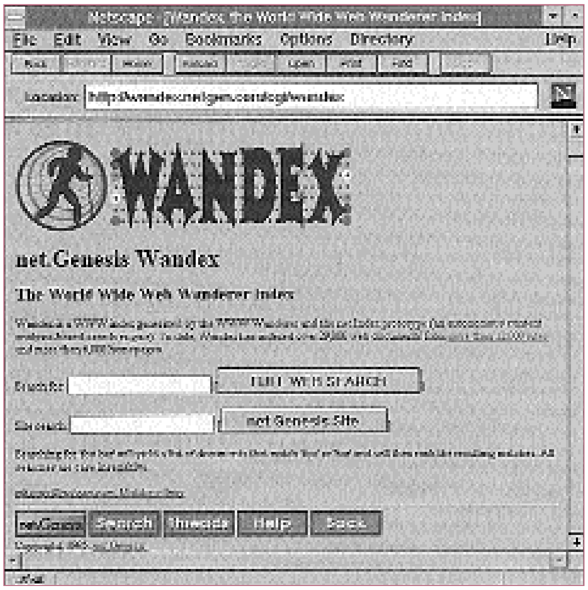
From its inception in 1969, when it consisted of four linked computers, until the 1990s, the Internet was largely used by the government, its contractors and the academic research community. As the Internet grew, creative programmers developed a range of Internet search tools including:

- Archie, a program to identify the software and text files available at the thousands of computers accessible by FTP (the

Internet's file transfer protocol). Currently there are about 38 Archie sites in existence. Since Archie searches file names and descriptions, files that lack descriptive names and appropriate descriptions can be missed. For Archie to work perfectly you must already know the name of the file you are searching for.

- Veronica, a program created to index information made available by gophers. Currently, about 20 Veronica "servers" exist. Veronica search results are frequently of limited relevance and the links to gopher menus provided are often incorrect, leading to missed connections. These errors occur because Veronica indexes of gopher sites are constructed only periodically; they cannot keep up with the constant changes that occur across the Internet.
- Jughead, a program that thoroughly indexes a single gopher site.
- Hytelnet, a program that organizes access to the many Internet-connected computer systems (telnet sites) and that permits users to sign on and search from remote locations.

Once developed, the Internet search systems tended to be overloaded with demand. In response, "mirror sites" that duplicated the original search sites were developed to help distribute this demand. For any given search system (Veronica, for instance) the different levels of activity due to varying time zones. Recommended Internet practice calls for searchers to use the search site that is geographically closest to them before attempting to use more remote search sites.


Netscape - [Wandex, the World Wide Web Wanderer Index]
File Edit View Go Bookmarks Options Directory Help
Location:
WANDEX
net.Genesis Wandex
The World Wide Web Wanderer Index
FULL WEB SEARCH
net.Genesis Site
Search
Threads
Help
Back

The search screen of Wandex,the World WideWeb Wanderer Index which may be found at http://wandex.net-gen.com/cgi/ wandex

What are the Problems with the Internet Search Tools?

Search results are often not precise. The Internet is in constant flux, and Archie, Veronica, and other indexes, however frequently generated, cannot keep up with that flux. Within the last few years, the ease of searching the Internet has improved dramatically, principally through the development (as part of the World Wide Web) of rapid search engines with relatively simple search statement or screen menu displays presented to users.

By 1993, gopher systems were being discovered and used by rapidly-rising numbers of new Internet explorers. But by the end of 1994, it was the World Wide Web, with its hypertext links, high-resolution graphics, and multimedia file content, that had taken over much of the Internet community's attention. At the same time, the Web incorporated the Internet searching tools that preceded it, making them easier to use.

Along with the availability of new Web-based Internet search tools, however, there is still a touch of the old chaos in that the tools are of widely varying power and coverage. They are appearing with little correlation at various sites; and the search results one gets from different tools that cover the same basic database (World Wide Web pages, for instance) are different. It is now possible to find **something** fairly easily and quickly on almost any imaginable topic; but, using any one of the new Web-based search tools, you generally cannot hope to have searched the Internet comprehensively on any such topic.

What Search Tools are Available on the World Wide Web for the Whole Internet?

As the Web has emerged as the most active and frequently-used Internet utility outside of email, it has also begun incorporating the other Internet features (gophers, newsgroups, FTP capability) that, up till now, have existed separately and involved the use of different software.

It is now possible, for instance, to do an Archie search from several Websites (see http://pubweb..nexor.co.uk/public/archie/servers.html), or to search thousands of USENET newsgroups for messages containing specific keywords. The latter can be done at a Website called the DejaNews Research Service (http://www.dejanews.com/), which as of the time of this writing is still available free of charge.

Another free service involving the USENET system is Stanford University's Netnews Filtering Service (http://woodstock.stanford. edu:2000/), through which one can submit an emailed profile of keywords that are run periodically-daily, if one prefers-against the newsgroups that are received at Stanford. Registrees are sent a regular email message consisting of the first several lines of each retrieved newsgroup message that matches the keyword profile. The full text of any partial messages that look interesting is obtainable on demand. Searches of “gopherspace” using public Veronica servers are possible through a Web browser at gopher sites like: gopher://gopher.scs.unr.edu/11 /veronica.

Other such services include a HYTELNET page (http://www.usask.ca/cgi-bin/hytelnet) that provides a rough index and access point to computers permitting remote access by the general public through the Internet telnet utility; the “WAIS Access through the Web” page (http://www.ai.mit.edu/the-net/wais.html) that lets some Web browsers access so-called Wide Area Information Server databases; and the “Search for Mailing Lists” page at the University of Indiana (http://SCWWW.ucs.Indiana.edu/mlarchive/). The latter is a database of email-based discussion groups running on the listserv, majordomo, or listproc software. It permits users to identify groups on their topics of interest from among nearly 13,000 available ones. Again, all of these rather amazing search services may be used free of charge as much as one likes.

There are also several powerful search services that concentrate on the World Wide Web itself. Among these are included Yahoo!, Lycos, the Open Text Web Index, InfoSeek, ALIWeb, the World Wide Web Worm, the World Wide Web Wanderer, WebCrawler, EINET Galaxy, Harvest Broker, CUI (Centre Universitaire d’Informatique), CUSI (Configurable Unified Search Index), and SavvySearch, among still others. Although several of these search systems started out as experimental systems developed in academic computer science departments, it is common for them to migrate from academia to a commercial site that has agreed to support the growing load of searches on its own hardware as a free public service.

How Do the Various WWW Search Tools Compare?

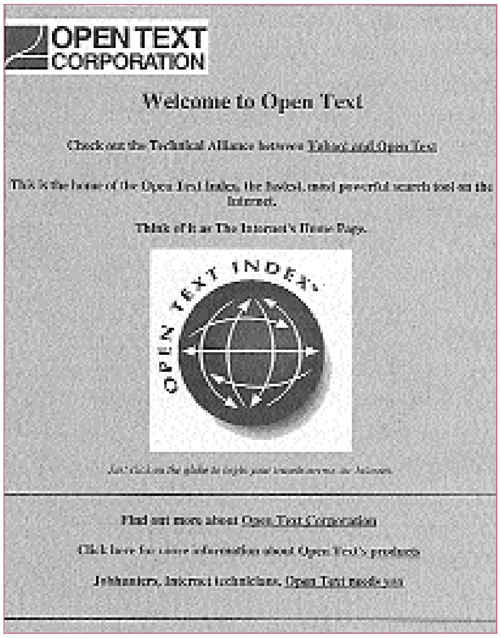
None of these search tools searches quite the same database in quite the same way. The syntax of search statements and the number of relevant items each system will allow you to display, for instance, vary widely. Some of the search services perform “relevance ranking” on search results to indicate how appropriate a particular retrieved item is to your search query. Each service offers a slightly variant range of search options and some are definitely easier and faster to use than others. Some are well-documented, some are not.

To illustrate this variability, the phrase “National Park Service” was used as the search statement on several of these systems in early September 1995. The results are in the box below.

Search Tool Name & URL	**Number of Items Found (Hits)**	**Comments**
Yahoo at http://www.yahoo.com	6 items	Retrieved 6 separate Web pages when the "National Park Service" was treated as a character string (adjacency searching) in which all words must appear next to each other in the same specified order before they will be retrieved.
Wandex, the World Wide Web Wanderer Index at http://wandex.netgen.com/cgi/wandex	9 items	
Open Text Index at http://www.opentext.com:8080/com.html	637 items	
WebCrawler at http://www.webcrawler.com	500 items displayed out of 1,466 retrieved	Note: Webcrawler only allows 500 retrievals, although it ranks them all in descending order of relevance.
Lycos at http://www.lycos.com/	974 items	Note: Lycos does not perform true adjacency searching, but instead simply finds all the selected terms in the material in whatever order. This search strategy increases hits but decreases relevance. Relevance ranking is performed on search results.
InfoSeek at http://www.infoseek.com/	>200 items	InfoSeek does not return a result greater than 200, and does not report how many additional relevant hits there are beyond 200.

What are the most Comprehensive Internet Search Tools on the WWW?

The most comprehensive databases of Web pages at this time (September 1995) appear to be Lycos, the Open Text Web Index, WebCrawler, and InfoSeek. The only commercial service among these, InfoSeek, may become more of a system of choice as the demand on the free systems grows, causing access problems and slowed response time, while its commercial nature limits use enough to avoid or delay overload. This is speculation, but it would be a logical outcome of the kind of growth in numbers of people with Internet access that has occurred over the last few years. For another site on Web searching tools, see "World Wide Web Robots, Wanderers and Spiders" at http://web.nexor.co.uk/mak/doc/robots/robots.html.



Welcome to Open Text

Check out the Technical Alliance between Yahoo! and Open Text

This is the home of the Open Text Index, the fastest, most powerful search tool on the Internet.

Think of it as The Internet's Home Page.

Just click on the globe to begin your search across the Internet.

Find out more about Open Text Corporation

Click here for more information about Open Text's products

Jobhunters, Internet technicians, Open Text needs you

The search screen of Open Text Index, which may be found at http://www. open-text.com: 8080/com.html

In a recent article ("Searching the World Wide Web: Lycos, WebCrawler and More," *Online*, July/August 1995, pp. 48-53), Greg Notess suggested a strategy that takes into account some of the differences and limitations of the Web search engines. "For single keyword searches of a large database, use Lycos. For multiword searches with an AND, try WebCrawler. For gopher resources, try veronica. And for a time-consuming comprehensive search, use CUSI" (p. 53). The CUSI site may be found at URL: http://web.nexor.co.uk/susi/cusi.html.

To these guidelines we might add that Yahoo! is especially useful if you want a fairly quick search through major Web resources only and when you want good search results without the requirement that they be comprehensive. Another reason for using Yahoo! is that it has extensive links to the other Internet searching systems and so has become an excellent gateway for those who want to run the same basic search through several different engines.

How Can Searching Techniques Be Improved?

Several factors account for differences in results from one search tool to another. The number of actual Web pages being searched by the systems varies. Some search engines bare also searching gopher and FTP sites as well as Web pages. Some services search the entire text of the Web pages they index, rather than just the title and perhaps some summary descriptive text. Some search engines do not perform true adjacency searching but rather just look for the appearance of all the search terms in whatever order-an approach that tends to raise the number of retrieved items.

Perhaps more importantly, some of these search tools and services explain the rules under which they operate better than others. It is not unusual for users to be presented with a search form, into which they are to type their search terms, with no up-front, explicit statement of whether the system will assume the words must be adjacent (such as National Park Service); or that they are related by an implied Boolean AND or OR operator ("architecture" and "vernacular" but not "garage"); or whether the system performs truncation by searching the words as character strings that may be separate words or the root characters of longer words (for example, "photo" which might get you "photograph" or "photosynthesis").

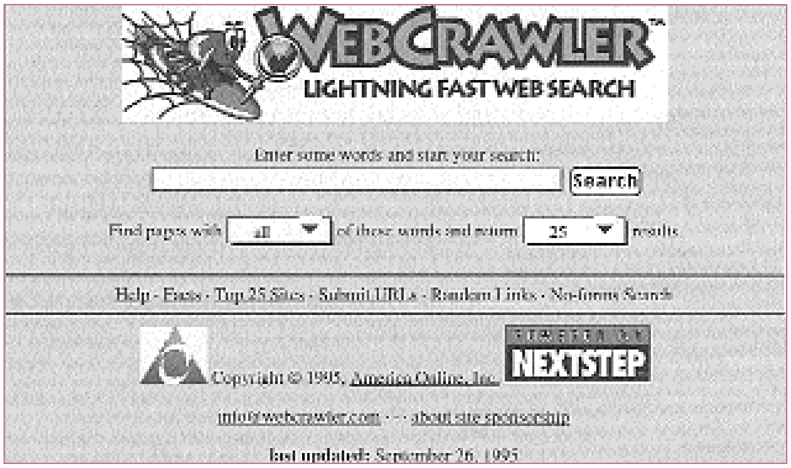
The Web search systems' main pages will usually have a link to explanatory text that will go into these details, or may perhaps link to a more elaborate version of the search form with all search features and options made explicit. But the casual user in a hurry may not bother with this and may therefore not be made aware of the ways to make a search more efficient. Retrieval on the WWW is currently more art than science. If comprehensiveness is your goal, carry out your search on several systems. A WWW search should always be considered more advisory than definitive.

How is News and Current Information Searched on the Internet?

Another important area for Internet search and retrieval is that of news and newsfeeds. It has become apparent to many news agencies that distribution of news stories and wire service material is yet another useful application of the Internet. There is, in fact, an entire news service (ClariNet eNews-see http://www.clarinet.com/) that is available only through the Internet.

One particularly ingenious (and, once again, free of charge) site on the World Wide Web is the CRAYON site. CRAYON here stands for "Create Your Own Newspaper." It is the product of Jeff Boulter, a student at Bucknell University, who has hyperlinked a variety of Internet news sources and enabled the user to design his or her own "front end" for selecting which of these sources to use and what order to view them in. The user designs a personal newspaper with its own title that will present the day's news whenever accessed.

The Crayon news options are so wide as to allow for true personalization of the result. It is


WEBCRAWLER™
LIGHTNING FAST WEB SEARCH
Search
Help · Facts · Top 25 Sites · Submit URLs · Random Links · No-forms Search
POWERED BY NEXTSTEP

Enter some words and start your search:

Find pages with all of these words and return 25 results

Copyright © 1995, America Online, Inc.

info@webcrawler.com · · · about site sponsorship

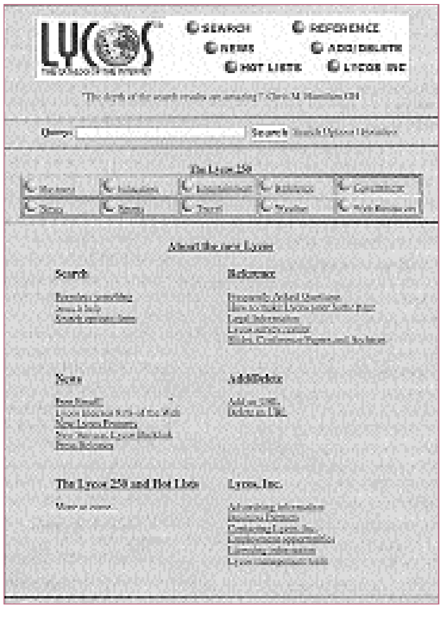
last updated: September 26, 1995

The search screen of WebCrawler which can be found at http://www.webcrawler.com/

possible, for instance, to include or exclude international news, sports, weather, and even comic strips, and to lay out the selected sources in customized order. Some of the news sources, like *Time* articles, are themselves hyperlinked to related past articles from the same source-providing a dimension of historical searchability to your newspaper.

The developer of CRAYON linked a variety of free sources from different sites on the Web into a cleverly-conceived central home page. This is a great example of the way originally disparate Web content can be "sliced and diced" and otherwise reassembled into new, value-added products through the device of hyperlinking. CRAYON, which had 81,403 subscribers internationally as of September 10, 1995, may be found at http://sun.bucknell.edu/~boultter/crayon/.

There are several sites, usually commercial, that now offer searching of general and specialized newsfeeds. Sometimes it is possible to set up an article profile that can retrieve news regularly on topics of continuing interest to the user. Sites to investigate for news retrieval include InfoSeek (http://www.infoseek.com/Home), NewsPage (http://www.newspage.com/), NewsHound, a San Jose Mercury News service (http://www.sjmercury.com/hound.html) and introNEWS at http://www.gold.net/info.highway/intemews/.



The search screen of Lycos which can be found at http://www.Lycos.com/

What is ENEWS and Where Do You Find It?

The field of automated news services, "enews", is a subset of electronic publishing (which includes magazine and journal publishing as well) and is evolving very rapidly right now. There are a number of Internet sites that attempt to offer comprehensive links to all those newspapers and other news publications (numbering in the hundreds) with some kind of Internet presence. That presence varies widely, from titles that offer a growing and searchable full-text backfile to those that simply show you part of a current issue and tell you how to subscribe to the paper edition.

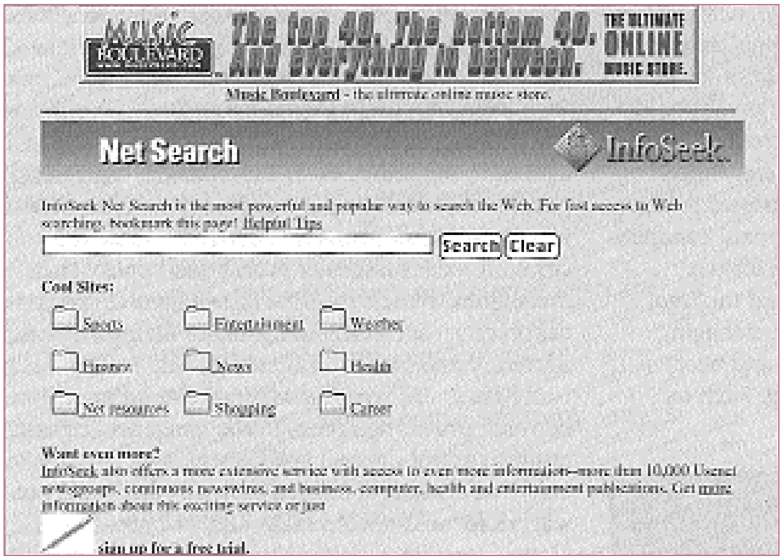
Some of the best sites that collect links to the nation's and the world's growing stock of enews titles include Taxi's International News (http://www.deltanet.com/users/taxicat/newstand.html), Steve Outing's Online Newspaper Services Resource Directory (http://www.nyc.pipeline.com/edpub/ e-papers.home.page.html), the Ecola EZ Home Page (http://www.ecola.com/ez/), UNCG's News and Newspapers Online Worldwide (http://www.uncg.edu/~cecarr/news/) and "Onramp Access-Newspapers on the Net" (http://www.onr.com/newspaper.html). These are only five of many such sites, some of them consisting of articles and general information on the phenomenon of electronic news, that are available through the World Wide Web.

Sites that provide comparable links and information on non-newspaper publications on the Internet include: "On-Line Magazines" (http://www.middlebury.edu/~otisg/zines.shtml), the ZineRak (http://www.greencart.com/zinerak/), John Labovitz's E-Zine List (http://www.mer.net/~johnl/ e-zine-list/index.html) and Internet Resources for Zines and EZines (http://www.acns.nwu.edu/ezines/ net-resources.html).

What Comprehensive Searching Services Exist?

There are a few Websites that offer more comprehensive searching capabilities. The All-In-One Internet Search site (http://www.albany.net/~wcross/all1gen.html) is actually a system that links to over 100 Internet databases. Through various points wit hin All-In-One, for example, one can search for the meanings of acronyms, perform a news search through CNN Interactive, determine the distance in miles between two user-specified U.S. locations, search the U.S. Code or the Code of Federal Regulations, or check out the Internet Movie Database. All-In-One may be expected to keep growing into a virtual reference collection as more database systems are linked to the All-In-One shell.

A fairly new commercial service that claims powerful search capabilities is the NLightN service (http://www.nlightn.com/), from The Library Company. NLightN searches the Internet as well as bibliographic databases and a set of full-text reference sources. It seems intent on becoming a single source through which the classic databases


Music Boulevard
The top 40. The bottom 40. And everything in between.
THE ULTIMATE ONLINE MUSIC STORE.

Music Boulevard - the ultimate online music store.

Net Search

InfoSeek

InfoSeek Net Search is the most powerful and popular way to search the Web. For fast access to Web searching, bookmark this page! Helpful Tips

Search Clear

Cool Sites:

Sports Entertainment Weather

Finance News Health

Net resources Shopping Career

Want even more?

InfoSeek also offers a more extensive service with access to even more information—more than 10,000 Usenet newsgroups, continuous newswires, and business, computer, health and entertainment publications. Get more information about this exciting service or just

sign up for a free trial.

The search screen of InfoSeek which can be found at http://infoseek.com/

well-known in libraries are accessible along with all the new content offered by the Internet.

What are Hotlists?

Finally, one must mention the large number of often subject-specific personal "hotlists" that have been brought up as Web pages. Many individuals have created and maintained guides to the Internet sites that cover topics of interest to them. As Web pages, these incorporate direct hyperlinks (mouse-clickable immediate transfers) to those sites. Such pro-bono enthusiasm means that t he individuals will probably maintain and build those sites over time through their own diligent research-saving all of us a lot of effort.

If you can find a person or agency that cares enough about a particular subject to maintain a home page for it, chances are that person or agency will rigorously maintain the page. Such sites can sometimes be found by doing a very broad term search on the Yahoo! site or any of the other search systems to which it provides links.

One can also consult the root index called the Wide Web Virtual Library (http://www.w3.org/hypertext/DataSources/bySubject/Overview.html) which organizes and links Websites as if they were part of some vast online research library-which, in fact, is just what they are.

In Conclusion

The Internet will continue to grow in content and complexity. Web pages grow over time, and incorporate new relevant links as they are discovered or as the developers of new pages have them linked to established ones. This is the almost self-indexing aspect of the Web. The search tools, indexes and resource lists that we have covered in this article are among those that can help keep one's Internet time as focused and productive as it is possible to be in such a dynamically evolving situation.

Hugh O'Connor is the Director of the Research Information Center of the American Association of Retired Persons.

Melissa Smith Levine

Electronic Publishing: A Legal and Practical Primer

Jules Veme's 1863 prediction that in the 20th century, "Photo-telegraphy allowed any writing, signature or illustration to be sent far away... Every house was wired," was a foresight worthy of the best clairvoyant.

The Opportunity

Recent technological advances provide astonishing opportunities to make information quickly, inexpensively, and widely available in ways never before possible. In order to minimize legal risk while using this new medium, consider developing a systematic electronic publishing policy that covers the technical and legal issues presented in this article. The policy must be free of confusing technical and legal jargon if it is to serve as a useful guideline for staff. Please note that this article presents a very brief overview of volatile and complex areas of the law.

The Challenge

Electronic publication over commercial online services and the Internet presents new legal challenges as legislatures, courts, and businesses catch up to technology already a part of everyday